POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates

Asier Uribeetxebarria[1] · José A. Martínez-Casasnovas[2] · Alexandre Escolà[1] · Joan R. Rosell-Polo[1] · Jaume Arnó[1]

Asier Uribeetxebarria (🖂)
Uribeetxebarria.asier@gmail.com

[1] Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, University of Lleida - Agrotecnio Centre, Rovira Roure, 191, Lleida, 25198, Catalonia, Spain

[2] Research Group in AgroICT & Precision Agriculture, Department of Environmental and Soil Sciences, University of Lleida - Agrotecnio Centre, Rovira Roure, 191, Lleida, 25198, Catalonia, Spain

Abstract Estimation of yield or other fruit quality parameter is of great interest to farmers to decide on management actions just before harvesting and, in any case, to anticipate and plan harvesting operations. Making accurate and reliable estimates often requires systematic sampling that, when covering the whole plot, can result in the use of a large number of samples and a significant effort in time and cost for fruit growers. Faced with this whole area sampling strategy, simple random sampling (SRS) using reduced sample sizes is currently a widely used technique despite the less precise estimates that it provides. In this work, different stratified sampling schemes have been tested to estimate yield (kg/tree), fruit firmness (kg/cm^2) and the refractometric index (ºBaumé) in a peach orchard located in Gimenells (Lleida, Catalonia, Spain). In contrast to SRS, the use of ancillary information (NDVI and apparent electrical conductivity, ECa) allowed sampling units or trees to be stratified according to two or three classes (strata) within the plot. The classes or homogeneous stratification zones were delimited by cluster analysis using, either separately or in combination, a multispectral airborne image (NDVI) and a ECa survey map acquired by means of a soil resistivity sensor (Veris 3100). Sampling schemes were then compared in terms of efficiency. In general, stratified sampling showed better results than SRS. Regarding yield estimates, stratified sampling according to two strata of NDVI allowed the sample size to be reduced by 17% compared to the SRS for the same precision. On the other hand, quality parameters may require different stratification strategies concerning the number of strata to be used. While ºBaumé was better-estimated using also stratified samples based on two strata of NDVI, fruit firmness showed better results when stratifying by three classes or strata of NDVI. In any case, neither the ECa nor the combined use of NDVI + ECa have improved sampling efficiency when used as ancillary maps for stratification.

Keywords Sampling efficiency · Fruit yield and quality · Peach · NDVI · Apparent electrical conductivity

Introduction

Sampling to estimate yield and/or fruit quality at harvest time is of great interest in fruit growing. However, reliable prediction of these parameters is not easy, especially when systematic sampling is usually replaced by a less complex simple random sampling (SRS) to reduce time and cost. In other occasions, random sampling raises doubts to both growers and advisors about how many trees should be sampled and, above all, which specific ones should be sampled within a plot. Facing this situation, there is a need to develop new and more precise methods with acceptable costs to guide fruit growers during field sampling. SRS is a widely used design, because it is relatively simple to implement by random selection of sampling units (trees) within the plot. However, SRS is inefficient when estimating parameters that show spatial autocorrelation within the plots (Webster and Lark 2013). Taylor et al. (2005) and Kazmierski et al. (2011) showed that vineyards are spatially variable and that grape yield usually follows well-defined and consistent spatial patterns over time. This same situation can be expected in fruit orchards and, for this reason, sampling methods that take into account the different areas within the plot with different expected yield values would be preferable to optimally locate sampling trees to obtain better yield estimates.

On the other hand, fruit growers can hire service companies that provide crop vigour and/or soil apparent electrical conductivity (ECa) maps obtained with suitable sensors (proximal and remote sensing). Normalized difference vegetation index (NDVI) derived from airborne images were used by Meyers and Vanden Heuvel (2014) to optimize sampling protocols in vineyard and reduce sample sizes. Applying a heuristic optimization algorithm (Tabu Search Algorithm) to NDVI images, specific samples to conform the spatial distribution of NDVI within the plot can be established (Meyers and Vanden Heuvel 2014). As NDVI is related to vine vigour, the method is a way for distributing sampling units by covering the areas of different vigour to capture vineyard canopy variability within the plot. This idea is also behind the method proposed by Carrillo et al. (2016) to improve grape yield estimates. The authors concluded with the need to consider a two-step sampling method combining NDVI-based samples with random vine samples to predict specific components of the productive potential in a vineyard. Regarding apparent electrical conductivity (ECa), there are several studies that address the use of ECa classified maps for site-specific management practices (Moral et al. 2010; Peralta and Costa 2013). The suitability of this information in fruit-growing sampling is a pending issue, although soil characteristics are expected to impact yield and/or quality parameters.

There are few studies on sampling in fruit orchards. Monestiez et al. (1990) proposed a geostatistical approach to assess spatial dependence between fruits to choose the most appropriate sampling designs inside the tree structure. Multilevel systematic sampling can also be an interesting option to estimate the number of fruits for yield forecasts (Wulfsohn et al. 2012), obtaining error coefficients of only 10%. More recently, sampling stratification using NDVI-based aerial images allowed different areas to be better delimited for sampling in nectarine orchards (Miranda et al. 2015), with a significant reduction in sample size (20-35%) compared to random sampling (Miranda et al. 2018). As is known, SRS can produce local clusters of trees and leave unrepresented areas within a plot (Webster and Lark 2013). Alternatively, farmers can consider using NDVI or ECa data to stratify samples, assuming that yield and quality parameters in orchards often present spatial autocorrelation and, what is more important, possible spatial cross-correlation with ancillary variables supplied by

POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

proximal and remote sensors of increasingly common use in agriculture. Cross-correlogram is a powerful tool to test the spatial correlation between two variables, and checking this spatial correlation may be the key factor before stratifying the samples.

The aim of this study was to investigate how the use of ancillary data (NDVI and ECa) in stratified sampling schemes can improve sampling efficiency compared to a SRS of equal size for the whole of a plot. Efforts in time and cost could be reduced with this new sampling strategy by optimizing sample sizes through the application of technological advances in the framework of precision agriculture. Sampling in orchards is then proposed as a design-based sampling strategy, making use of classical sampling theory (that is, assuming normality and independence of observations). This may be a limitation in plots with spatial autocorrelation. However, the use of geostatistical methods is beyond the scope of this paper.

Materials and methods

Study plot

The research was conducted in a peach orchard (*Prunus persica* cv. 'Platycarpa') located at the IRTA Experimental Station (41° 39' 19" N, 0° 23' 36" E, ETRS89) in Gimenells (Lleida, Catalonia, Spain). The plot covered an area of 0.65 ha, and was planted in 2011 according to a 5 x 2.80 m pattern (Fig. 1). Soil was classified as Petrocalcic Calcixerept (Soil Survey Staff 2014), and it was a well-drained soil without salinity problems. The presence of a petrocalcic horizon at a variable depth (0.4-0.8 m) and high $CaCO_3$ content were the main soil limiting factors. The horizon may be at shallow depth due to successive earth movements and tillage operations that, over time and since 1946, have contributed to modify in shape and size of the parcelling in the farm. The climate is typical of hot semi-arid areas, with strong seasonal temperature variations (cold winters and hot summers). Annual precipitation is frequently below 400 mm, and basically distributed from September to May.

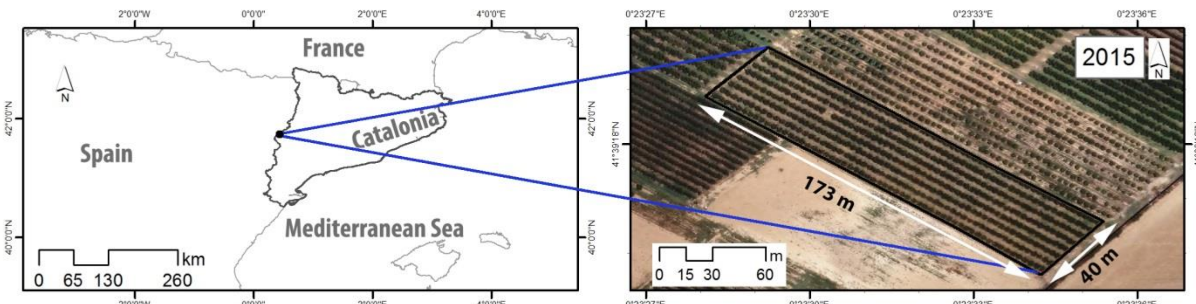

Fig. 1 Location of the study area (left), and orthophoto of the peach orchard plot in 2015 (right).

Sample size and stratification

Three production and quality variables were sampled within the plot: yield (kg/tree), fruit firmness (kg/cm^2) and refractometric index (ºBaumé). To determine the sample size, an aerial multi-spectral image was taken on June 9th, 2015. The image resolution was 0.25 m/pixel. Once the canopies were individually delimited on the basis of this image (ESRI® ArcMap^{TM}

POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

10.4.1) to obtain a map of georeferenced trees within the plot (statistical population), a weighted average value of NDVI according to the area of the canopy was assigned to each tree. These tree-averaged NDVI values were then used as base data for determining the sample size for a SRS without replacement using Eq. 1:

$$n = \frac{\zeta_{\alpha/2}^{2} CV^{2}}{E_{R}^{2}} \qquad (1)$$

where n is the sample size (number of trees) assuming sample independence, $\zeta_{\alpha/2}$ (1.96) is the value of the standard normal variate for a 95% confidence ($\alpha = 0.05$), CV is the Coefficient of Variation (17.5% in the present case), and E_R is the relative error assumed (10%). The result of Eq. 1 was 12 sampling trees that were first randomly distributed within the plot (sampling scheme A, Fig. 2). Apart from being a usual index for detecting spatial variability in tree crops (Kazmierski et al. 2011), the use of NDVI for this approach was justified because previous successful applications in fruit sampling were known (Miranda et al. 2015, 2018).

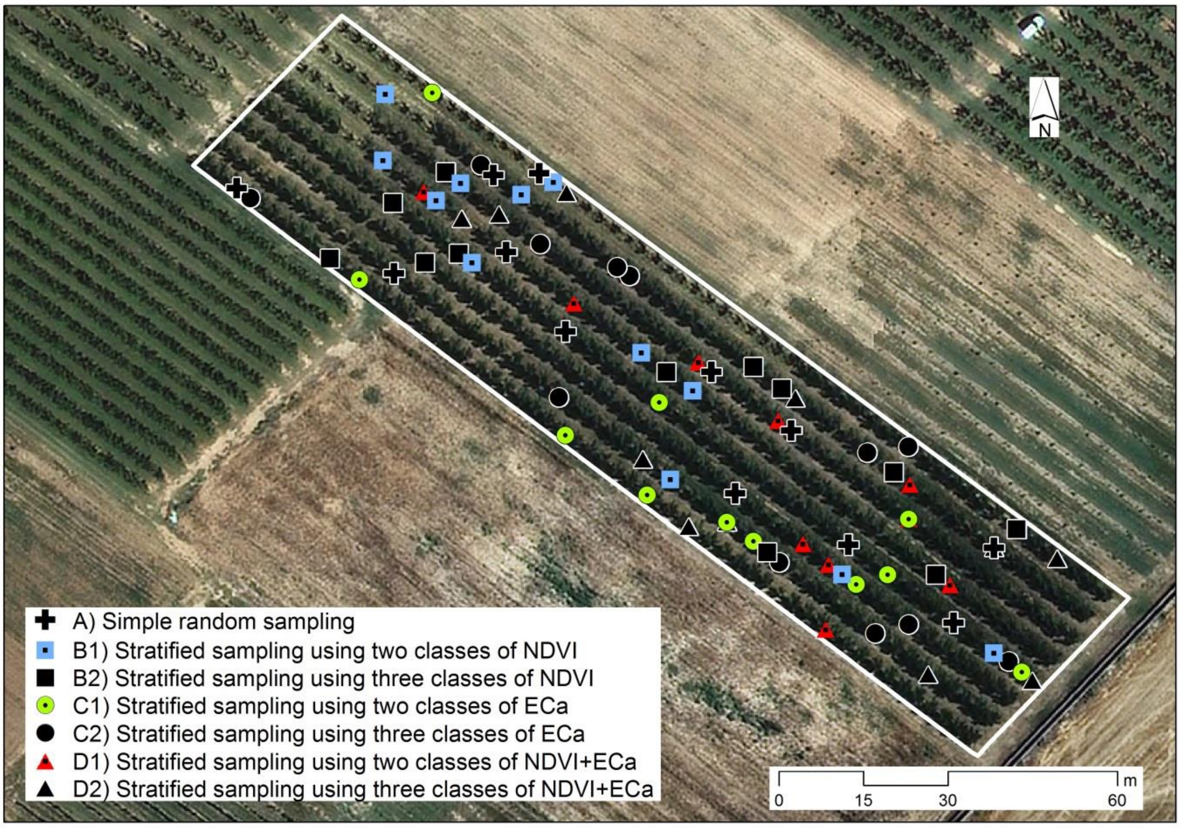

Fig. 2 Sampling units (trees) corresponding to seven different sampling schemes.

Additional schemes were tested in which new sampling trees (twelve in each case) were first obtained by stratified random sampling according to two and three classes of NDVI. Specifically, NDVI classified maps were built by clustering interpolated NDVI values (NDVI raster map) using the unsupervised classification algorithm ISODATA (Jensen 1996). The process on which this algorithm is based is well known. Assigning an arbitrary mean to each class, pixels were then successively reassigned minimizing the Euclidean distance from each

POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

pixel to the mean value of the class. Each iteration, class means were recalculated and pixels were reallocated until the last iteration is reached, or the number of pixels that change from one class to another does not exceed a certain threshold (Guastaferro et al. 2010). The same strategy (stratified sampling based on clustered maps) was repeated using the information provided by a Veris 3100 ECa surveyor. As a widely used sensor for soil characterization (Sudduth et al. 2005), the information provided may be very useful in sampling given the soil-tree interaction. This sensor measured the ECa at two soil depths: shallow (0-0.3 m) and deep (0-0.9 m). Both ECa value layers were interpolated by ordinary kriging, and ECa classes were established based on the cluster analysis of the two maps (shallow and deep) simultaneously. Finally, the same procedure was repeated again by taking all three ancillary layers (NDVI, shallow ECa and deep ECa). In short, seven sampling schemes (including scheme A) were compared to each other based on a total number of 84 sampled trees (7x12) within the plot (Fig. 2). Figure 3 shows five of the proposed sampling schemes, (i) SRS (scheme A), (ii) stratified sampling based on two classes of NDVI (scheme B1), (iii) stratified sampling based on three classes of NDVI (scheme B2), (iv) stratified sampling based on two classes of ECa (scheme C1), and (v) stratified sampling based on three classes of ECa (scheme C2). Schemes that use both information layers (schemes D1 and D2) are not shown. In each case, sampling trees within each stratum were randomly sampled without replacement.

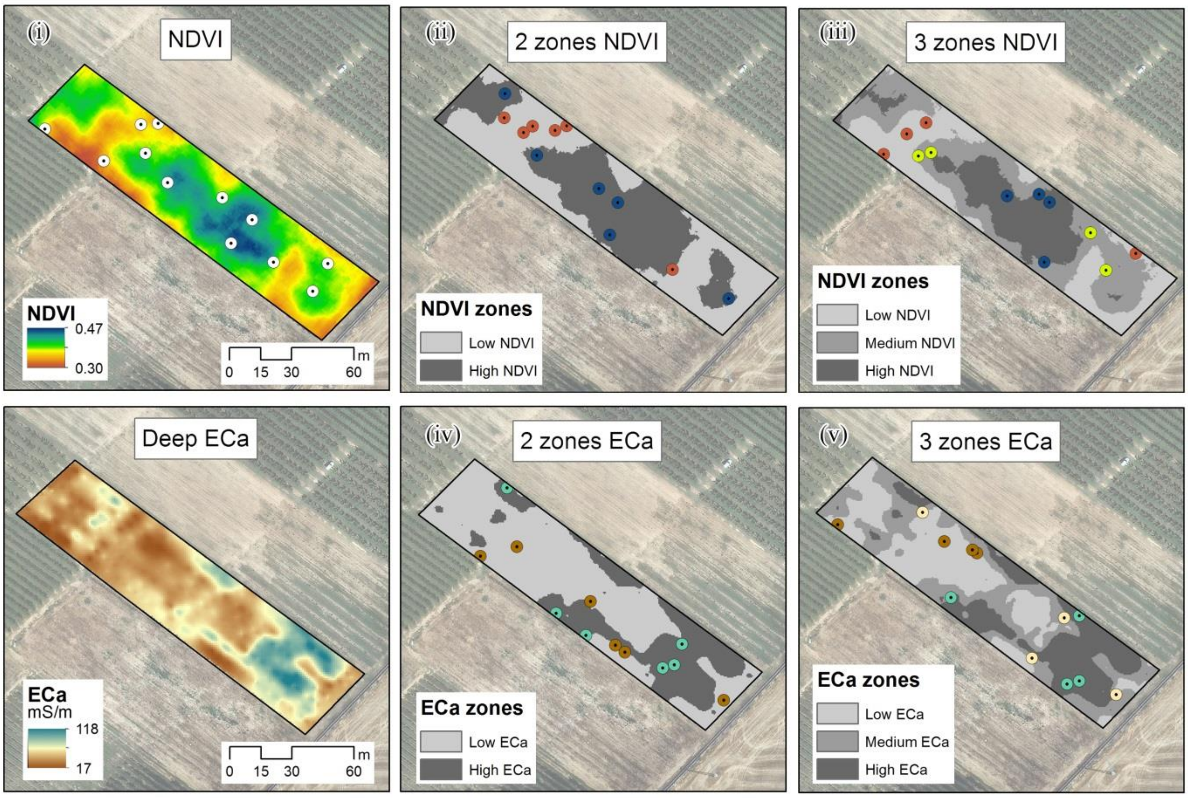

Fig. 3 Sampling schemes: (i) simple random sampling, (ii) stratified sampling by NDVI (two strata), (iii) stratified sampling by NDVI (three strata), (iv) stratified sampling by ECa (two strata), (v) stratified sampling by ECa (three strata).

Estimation in stratified sampling schemes

In a SRS approach, the sample mean ($\overline{z}_{SRS}$) has proven to be an unbiased estimator of the population mean (μ), with a variance that can be estimated by $v(\overline{z}_{SRS}) = \frac{s^2}{n} \times (1-f) = \frac{s^2}{n} \times \left(1-\frac{n}{N}\right)$ where s^2 is the sample variance, and $1-f = 1-\frac{n}{N}$ is the finite population correction or fpc, where *n* is the sample size and *N* the size of the population (459 trees in the plot under study). As the interest was to work with small size samples, confidence limits for the mean can be formulated as $\overline{z}_{SRS} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{1-f}$, where $\overline{z}_{SRS}$ is the sample mean, $\frac{s}{\sqrt{n}}\sqrt{1-f}$ is the standard error of the mean, and $t_{\alpha/2}$ is the Student's *t* value corresponding to *n*-1 degrees of freedom for a 95% confidence.

In order to sample more efficiently, other sampling schemes were used by stratifying the 12 sampling trees according to two strata (6 trees per stratum) or three strata (4 trees per stratum). As a reminder, strata corresponded to the classes obtained after classification of the plot according to NDVI, ECa or both auxiliary data layers. The different stratifications produced classes that were not equal in area (therefore, with different number of trees per stratum), and so the plot mean (μ) was then estimated for *K* classes (strata) within the plot using a weighted average as suggested by Cochran (1977), and more recently by Webster and Lark (2013) in what is called regional classification techniques:

$$\overline{z}_{StRS} = \sum_{k=1}^{K} w_k \times \overline{z}_k \tag{2}$$

where $\overline{z}_k$ is the sample mean of the *k*th class, and w_k allowed the number of individuals (trees) of the *k*th class to be weighted using Eq. 3,

$$w_k = \frac{N_k}{N} \tag{3}$$

where N_k is the number of trees within stratum *k*, and *N* is the total number within the plot.

As in SRS, confidence limits were obtained using the standard error of the mean, in this case, the square root of the estimated variance (Cochran, 1977):

$$v(\overline{z}_{StRS}) = \sum_{k=1}^{K} \frac{w_k^2 s_k^2}{n_k} \times (1-f_k) \tag{4}$$

where s_k^2 is the within-class sample variance of the *k*th stratum, n_k is the sampling trees within the stratum (6 or 4), and $1-f_k$ is the fpc for the *k*th stratum calculated as $1-f_k = 1-\frac{n_k}{N_k}$. Finally, the value $t_{\alpha/2}$ was adjusted for each stratified sampling scheme according to an effective number of degrees of freedom as established by Cochran (1977) in these cases.

The above confidence intervals were obtained assuming normality of observations. Since this hypothesis was not tested (for example, using Shapiro-Wilk test), additional intervals were calculated by applying a bootstrap estimation with the aim of contrasting the results. Bootstrap is a method of resampling to obtain approximately the precision of an estimator without hypothesizing about its distribution. Thus, for each set of 12 sampling trees corresponding to the different sampling schemes (which are now the statistical population), sampling is done with replacement until obtaining 1000 sample arrangements each of equal size 12. By averaging the 12 values in each new sample, it is known that 1-α level confidence intervals can be obtained from the distribution of the 1000 calculated mean values through the use of the percentile method (Efron 1982). Specifically, confidence limits were established excluding the 1000 $\alpha/2$ values located at the extreme positions of the distribution (α = 0.05). In all cases, sample arrangement generation was performed by programming in R software, version 3.3.2.

Sampling efficiency

The most interesting sampling scheme is that which provides, on average, the least mean squared error (*MSE*). Since the seven sample means were unbiased estimators of the plot mean, *MSE* can be used as a measure of accuracy. Coinciding *MSE* with the variance (Eq. 5), efficiency to estimate the plot or population mean (μ) can be established as the inverse of the estimated variance of the sample mean.

$$MSE(\overline{z}) = [bias(\overline{z})]^2 + v(\overline{z}) = v(\overline{z}) \tag{5}$$

To compare any of the stratified sampling schemes ($\overline{z}_{StRS}$) with respect to the simple random sampling design ($\overline{z}_{SRS}$), the relative efficiency (*RE*) was obtained as shown in Eq. 6:

$$RE = \frac{Efficiency\ (\overline{z}_{StRS})}{Efficiency\ (\overline{z}_{SRS})} = \frac{v(\overline{z}_{SRS})}{v(\overline{z}_{StRS})} \tag{6}$$

where $v(\overline{z}_{StRS})$ was in each case the variance of the stratified sample mean, and $v(\overline{z}_{SRS})$ or variance of a random sample mean of the same size (taken as reference) was best estimated by applying the method suggested by Cochran (1977). Specifically, given the results of a stratified random sample, an unbiased estimator of the variance of the mean for a simple random sample from the same population is (Eq. 7),

$$v(\overline{z}_{SRS}) = \frac{(N-n)}{n(N-1)}\left[\frac{1}{N}\sum_k^K \frac{N_k}{n_k}\sum_i^{n_k} z_{ki}^2 - \overline{z}_{StRS}^2 + v(\overline{z}_{StRS})\right] \tag{7}$$

where z_{ki} were the values sampled at trees within stratum *k*. The other parameters are those stated in previous paragraphs. By averaging the six previously calculated variances (one for each stratified sample) with the variance previously obtained for scheme A (SRS), the resulting variance, $v(\overline{z}_{SRS})$, was the one used in the calculation of the *RE*. The reason for using Eq. 7 was the use of non-proportional allocation of sampling trees, that is, the same number of sampling units (6 or 4) was assigned regardless of the size (number of trees) of each stratum.

Both the *MSE* and the *RE* were the statistics that served for the comparison of the different sampling schemes and, above all, for the verification of the possible gain due to stratification. Knowing the *RE* allowed the necessary sample size for the same precision to be compared between sampling schemes. Low *MSE* values and values of *RE* greater than 1 are those sought for stratified sampling schemes.

An estimate of the population mean

Considering the spatial distribution of the 84 sampling trees resulting from the seven sampling schemes (7x12) (Fig. 2), it is important to emphasize that only 5% of the plot area resulted in a weak sampling density, i. e. with sampling units separated from each other by a distance larger than 9.78 m (range of the NDVI exponential variogram, not shown). So, the sampled information contained in these 84 trees was finally considered to estimate the mean of the plot as accurately as possible by calculating a weighted average of the means of the samples. The most accurate linear combination of the seven independent sample means was obtained by assigning proportionally greater weighting to the more precise (Eq. 8),

$$\bar{z}_w = \sum a_i \times \bar{z}_i \qquad (8)$$

where $\bar{z}_w$ is the weighted average for the plot, $\bar{z}_i$ are the sample means calculated for each of the seven sampling schemes, and a_i are the relative weights calculated using the inverse of the variance of the sample means (Eq. 9):

$$a_i = \frac{1/v(\bar{z}_i)}{\sum 1/v(\bar{z}_i)} \qquad (9)$$

Goodness of stratification

Finally, and as already mentioned, stratified sampling schemes were based on a previous classification of the plot. A more accurate and efficient estimation of the mean was linked to the ability of the NDVI and/or ECa auxiliary layers to discriminate different mean values between classes, while the values within the classes have lower intra-class variability compared to the total variability of the plot. A parameter that served to judge the goodness of these classifications was the relative variance ($r_V = s_W^2/s_T^2$), where s_W^2 was the pooled or average within-class variance, and s_T^2 was the total variance in the sample (Webster and Lark 2013). Used in the form of its complement (Eq. 10),

$$1 - (s_W^2/s_T^2) = 1 - r_V \qquad (10)$$

it allowed values close to 1 to be obtained for those more effective sampling schemes. Values close to 0 or even negative corresponded to non-effective stratifications.

Spatial cross-correlation

To check stratified sampling results using ancillary variables, bivariate Moran's coefficient was also calculated to assess the spatial cross-correlation between ancillary information layers (NDVI and ECa) and the sampled yield and quality variables (GeoDa 1.12 software, Anselin et al. 2010). Hypothetically, the most efficient stratified sampling schemes would be those with significant spatial correlation with the variables to be sampled. Having verified significant spatial autocorrelation for the three variables of interest (Moran's *I* coefficient on the total of 84 sampled trees, data not shown), assessing spatial cross-correlation between ancillary variables and sampled variables could report information (even if a posteriori) on what ancillary information was most convenient in each case. However, it must be said that the use of geostatistical methods was beyond the scope of this paper. So, classical sampling theory was prevalent to assess stratified methods in this work under what is called design-based sampling strategies (Brus & de Gruijter 1997).

Results and discussion

Table 1 shows the mean squared error (*MSE*) and the relative efficiency (*RE*) for the different sampling schemes tested. For each of the variables (yield, fruit firmness and refractometric index), confidence intervals (CIs) for the population mean (μ) are also shown. Two types of confidence intervals were built for each sampling scheme as a result of using, i) the standard error of the corresponding sample mean (parametric approach) or ii) the non-parametric bootstrap approach. In the same Table 1, the weighted average of the plot $\bar{z}_w$ for each field variable is added next to the sample means. By completing this table of results, each sampling scheme is valued according to the goodness of stratification using the value 1 minus the relative variance.

Concerning the confidence intervals for the mean, bootstrap CIs were always slightly narrower compared to CIs based on the normality of the sample means. This may be due to the asymptotic approximation of the bootstrap method and, in any case, could prove the non-normality of the distributions. However, and for comparison purposes, relative efficiency (*RE*, Table 1) based on the estimated variances of the sample means (Eq. 6) was the statistic taken as a reference instead of the CIs. As general results, stratified sampling seemed to improve efficiency (*RE*) compared to SRS, mostly for the quality variables (fruit firmness and refractometric index). The improvement in yield estimation efficiency using stratified sampling was lower than for quality variables, and it was only evident in very particular cases of stratification. To aid interpretation, a more detailed analysis of the results in Table 1 is addressed in the following sections.

Sampling to estimate yield

Compared to the other sampling schemes, stratified sampling based on two classes of NDVI (scheme B1) was the one that showed the best results in estimating yield, with an expected average error $\left(\sqrt{MSE}\right)$ of 2.71 kg/tree (Table 1). Surprisingly, when stratifying the sample in three NDVI classes (scheme B2), the method failed to improve the efficiency or precision compared to SRS. This result could be explained by the poor effectiveness of the stratification (negative value of $1 - r_V$). In fact, negative values of the goodness of the stratification have always been obtained in those inefficient schemes with *RE* less than 1.

Concerning the use of ECa as ancillary information, stratifying the sample according to three classes (strata) of soil conductivity (scheme C2) has also shown better efficiency results than SRS. However, the error (*MSE*) and relative efficiency (*RE*) are not as good as in scheme B1 (stratification according to two classes of NDVI). Again and unexpectedly, the stratified sampling has shown better efficiency than SRS despite the poor result of the goodness of the stratification (positive but very low value of $1 - r_V$, and very far from the optimal values close to 1).

The choice between using scheme B1 (stratifying by using the NDVI) and scheme C2 (stratifying by using the ECa) is not easy. An analysis of the special characteristics of the plot can help to understand the sampling results for later decision-making. In the plot under study, affected by the presence of a petrocalcic horizon and high $CaCO_3$ content, some advantage was expected by stratifying the sample using a classified map of the ECa. In fact, significant inverse spatial cross-correlation was obtained between yield and ECa using the bivariate Moran's I_B statistic (Table 2). As high $CaCO_3$ content is a limiting factor of yield, with high ECa values usually associated with low yields (Martínez-Casasnovas et al. 2012; Ortega-Blue and Molina-Roco 2016; Uribeetxebarria et al. 2018), the spatial variation of ECa could make it advisable to stratify on the basis of this information layer instead of using an NDVI map. However, given the also significant spatial correlation between NDVI and yield (Table 2), the best efficiency results, and the simplicity in managing the stratification in only two strata, the B1 scheme is the option to recommend. In fact, NDVI has been used successfully to guide sampling for yield forecasting tasks in many crops (Fortes et al. 2015; Miranda and Royo 2003; Taylor et al. 2010).

Table 1 Efficiency parameters for the sampling schemes tested.

Sampling scheme	Mean ($\bar{z}$)	$(MSE)^{1/2}$	CI_L	CI_U	CI_{LB}	CI_{UB}	*RE*	$1-r_V$
Yield (kg/tree)								
Weighted average of the plot	24.49							
A	26.36	2.32	21.26	31.47	22.41	30.73		0.00
B1	24.33	2.71	17.93	30.73	19.65	30.09	1.20	0.04
B2	24.29	3.65	15.37	33.21	18.28	30.58	0.66	-0.10
C1	23.57	3.70	14.83	32.31	16.90	29.43	0.64	-0.09
C2	24.58	2.82	17.90	31.26	19.11	30.20	1.10	0.07
D1	22.09	3.30	14.29	29.89	16.32	27.00	0.81	-0.08
D2	24.21	2.87	16.23	32.20	18.84	29.18	1.06	0.08
Fruit firmness (kg/cm^2)								
Weighted average of the plot	4.33							
A	4.10	0.30	3.44	4.76	3.51	4.60		0.00
B1	4.31	0.22	3.82	4.81	3.87	4.76	1.80	0.13
B2	4.26	0.17	3.88	4.65	3.89	4.67	3.09	0.28
C1	4.27	0.35	3.45	5.08	3.56	4.79	0.69	-0.08
C2	4.75	0.27	4.13	5.37	4.28	5.22	1.18	-0.19
D1	4.40	0.33	3.66	5.14	3.40	4.92	0.80	0.19
D2	4.28	0.29	3.36	5.21	3.71	4.83	1.01	0.12
Refractometric index (ºBaumé)								
Weighted average of the plot	6.86							
A	7.02	0.14	6.71	7.32	6.80	7.31		0.00
B1	6.55	0.10	6.32	6.78	6.34	6.74	3.77	0.10
B2	6.63	0.19	6.10	7.16	5.90	7.09	1.09	0.54
C1	7.22	0.17	6.81	7.63	6.99	7.54	1.40	-0.09
C2	6.88	0.10	6.64	7.12	6.70	7.12	4.04	0.21
D1	6.80	0.32	5.99	7.61	6.37	7.32	0.39	-0.05
D2	7.31	0.17	6.86	7.76	6.88	7.66	1.29	0.38

A (Simple random sampling); B1 and B2 (NDVI stratified sampling, 2 and 3 classes); C1 and C2 (ECa stratified sampling, 2 and 3 classes); D1 and D2 (combined NDVI + ECa stratified sampling, 2 and 3 classes). *MSE* (Mean Squared Error), CI_L and CI_U (lower and upper confidence interval considering normality), CI_{LB} and CI_{UB} (lower and upper CI using bootstrap), *RE* (relative efficiency), r_V (relative variance).

From a practical point of view, as scheme B1 was more efficient (*RE* out of 1.20, Table 1), a similar efficiency for the SRS (scheme A) could be reached using a smaller sample size, theoretically equal to *n*(SRS)/*RE* (12/1.20). In short, stratified sampling according to two strata of NDVI allowed the sample size to be reduced by 17% compared to the SRS for the same precision.

Table 2 Spatial cross-correlation between NDVI and ECa ancillary information layers and the sampled yield and quality variables

Ancillary information	Sampled fruit variable	Bivariate Moran's I_B coefficient*	Pseudo p-value**
NDVI	Yield	-0.147	0.011
NDVI	Fruit firmness	0.199	0.004
NDVI	Refractometric index	-0.215	0.002
ECa	Yield	-0.315	0.001
ECa	Fruit firmness	-0.059	0.173
ECa	Refractometric index	0.029	0.306

*Global spatial statistic to estimate the spatial cross-correlation between ancillary and sampled variables. Correlation calculated based on 84 sampling trees using GeoDa 1.12 software (Anselin et al. 2010).
**Significance test was based on 999 permutations to generate the reference distribution under the null hypothesis of spatial randomness. The observed statistic was then compared to this distribution to calculate a so-called pseudo p-value (0.001 is the most extreme pseudo p-value under this scenario).

Sampling to estimate fruit quality parameters

Stratified sampling schemes worked differently when estimating fruit quality parameters. Regarding fruit firmness (Table 1), scheme B2 was clearly better in both *MSE* and efficiency (*RE* greater than the other sampling schemes). A significant spatial cross-correlation between NDVI and firmness (the greater the NDVI, the greater the firmness) could explain this result (Table 2). Likewise, stratifying sampling trees based on three strata of NDVI allowed spatial classification in fruit firmness to be more effective ($1 - r_V$ = 0.28). Concerning the sugar content of the fruit (refractomeric index), the results were somewhat difficult to interpret. Again, NDVI correlated spatially in a significant way (Table 2), showing an inverse relationship. However, among the two proposed schemes (B1 and B2), it was the B1 scheme (stratification through 2 strata) that achieved the best efficiency (*RE* value of 3.77). On the other hand, this result was somewhat inconsistent with the effectiveness of the stratification. Sampling trees were optimally classified using three classes of NDVI as suggested by the goodness of stratification ($1 - r_V$ in Table 1). Therefore, there was a discrepancy in the efficiency scores between schemes B1 and B2, and reasonable doubts arise as to whether to use two or three strata to stratify the samples. This situation can occur because the sampling trees can be well segmented (reducing the average within-class variance) and, however, presenting a variance of the sample mean too high (poor value of the *RE*). Since the fundamental criterion sought is to increase the *RE*, the B1 scheme would be the recommended option in this case by making compatible the values of *RE* (Table 1) and spatial correlation (Table 2).

The relationship between NDVI and some quality parameters has been shown in other studies (Zude-Sasse et al. 2016; Martínez-Casasnovas et al. 2012). Many times, fruits achieve lower sugar content (ºBaumé) in the areas with the highest NDVI values. Inversely, vigorous canopies with high amount of leaves (and higher NDVI values) can shade the fruits affecting fruit ripening and, as happens in viticulture (Vanden Heuvel et al. 2002), producing greener fruits with higher firmness values. This would explain the significant spatial relationship between NDVI, fruit firmness and ºBaumé within the plot (Table 2).

Regarding the use of ECa as ancillary information to stratify the quality, specifically ºBaumé in fruit, the results have been contradictory. While sampling scheme C2 (three strata of ECa) has shown the highest relative efficiency (*RE* of 4.04) together with good goodness of stratification, spatial cross-correlation between both parameters (ECa and ºBaumé) was not significant (Table 2). Since spatial correlation is an essential requirement to justify the suitability of stratification, the use of ECa was not a priori an interesting option to stratify the sampling. Nevertheless, as already said before, there could be an opportunity to use it to efficiently estimate yield in this plot.

Lessons learned for future research

Estimated variance of the mean in stratified sampling (StRS) is usually expected to be less than the variance of a simple random sample (SRS) of the same size (Cochran 1977). Once the sample size is decided (in our case, 12 sampling units), variance for StRS is finally influenced by the particular allocation of the sampling units between the strata. Two or three strata were delimited in this work using auxiliary information maps (NDVI or ECa), to then allocate the same number of sampling units (trees) for all strata (6 or 4 trees per stratum if two or three strata were used, respectively). This procedure probably resulted in a non-optimal allocation of the sampled trees and, as a consequence, in a possible greater uncertainty (or less precision) of the estimates. Cochran (1977) managed to evaluate, for a fixed sample size *n*, the effect of the deviation from an optimal allocation of sampling units in stratified samples. According to this approach, no significant increase in the variance (or significant loss of efficiency) was expected due to having used the same number of trees per stratum (data not shown). The use of identical allocation in each stratum was for reasons of simplifying the whole process for the farmer. However, it would be advisable in future works to opt for the proportional allocation of sampling trees according to the size of the strata to possibly minimize the variance of the stratified means. Moving away from the optimal allocation of sampling trees should be especially sensitive in yield estimation. This would explain why the variance of the mean in the stratified sampling according to three strata of NDVI was unexpectedly greater than the variance of the SRS. Interesting results comparing proportional and optimum allocation can be found in Brus (1994).

Finally, being in agreement with other studies (Meyers et al. 2011), sample stratification making use of ancillary information is a possibility to take into account in fruit growing. Cluster analysis has been the option used in this work for the construction of strata. A pending issue for future work is to check other methods to optimize strata such as, for example, the well-known rule of the cumulative root of the frequency function (see Cochran 1977). Although both SRS and stratified sampling provide unbiased estimates of the population mean, stratifying the sample (Lark and Marchant 2009) is a way to (i) get more precise estimates (or estimates with less uncertainty), or (ii) reduce the sample size for a certain precision or efficiency. However, there is a major limiting factor as it is necessary for the ancillary information to be spatially correlated with the variable to be sampled. If this requirement is met, sample estimates can improve in precision. Ultimately, fruit growers and technical advisors can benefit from positive impacts on operating time and cost.

Conclusions

The use of ancillary data such as NDVI in stratified sampling schemes allows yield and quality parameters in a peach orchard to be estimated with greater precision (or greater efficiency). For fruit firmness, the stratification in three strata (scheme B2) is the most recommendable option, achieving almost triple the efficiency compared to simple random sampling (SRS). This means being able to reduce the sample size by almost 67% for the same precision of the estimates. On the other hand, refractometric index may require a simpler stratification scheme using only two NDVI classes (scheme B1). In terms of yield estimation, the 20% higher efficiency of also stratified sampling according to two strata of NDVI (scheme B1) allowed the sample size to be reduced by 17% compared to SRS. In no case the ECa or the combined use of NDVI and ECa have provided substantial advantages compared to the use of NDVI as a single layer of ancillary information. So, the recommendation is to use NDVI as ancillary information to more efficiently estimate yield and quality variables in peach. However, and especially for yield estimates, caution must be taken at the time of allocating sampling trees by strata.

Acknowledgements

This work was funded by the Spanish Ministry of Economy and Competitiveness through the project AgVANCE (AGL2013-48297-C2-2-R). The authors also thank the IRTA Experimental Station in Gimenells (Lleida, Spain) for the possibility of carrying out this sampling study on a peach orchard.

References

Anselin, L., Syabri, I., & Kho, Y. (2010). GeoDa: An Introduction to Spatial Data Analysis. In: M. M. Fischer, A. Getis (Ed.) Handbook of Applied Spatial Analysis, Berlin, Germany: Springer, pp. 73-89.

Brus, D. J. (1994). Improving design-based estimation of spatial means by soil map stratification. A case study of phosphate saturation. *Geoderma, 62*, 233-246.

Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma, 80*, 1-44.

Carrillo, E., Matese, A., Rousseau, J., & Tisseyre, B. (2016). Use of multi-spectral airborne imagery to improve yield sampling in viticulture. *Precision Agriculture*, *17*, 74-92.

Cochran, W. G. (1977). *Sampling Techniques*. New York, NY, USA: John Wiley & Sons, Inc. 428 pp.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. 85 pp.

Fortes, R., Prieto, M. H., García-Martín, A., Córdoba, A., Martínez, L., & Campillo, C. (2015). Using NDVI and guided sampling to develop yield prediction maps of processing tomato crop. *Spanish Journal of Agricultural Research, 13(1)*, e02-004, 9 pages.

Guastaferro, F., Castrignanò, A., De Benedetto, D., Sollitto, D., Troccoli, A., & Cafarelli, B. (2010). A comparison of different algorithms for the delineation of management zones. *Precision Agriculture*, *11*, 600-620.

POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

Jensen, J. R. (1996). *Introductory Digital Image Processing: A Remote Sensing Perspective*. Englewood Cliffs, NJ, USA: Prentice Hall. 316 pp.

Kazmierski, M., Glemas, P., Rousseau, J., & Tisseyre, B. (2011). Temporal stability of within-field patterns of NDVI in non-irrigated Mediterranean vineyards. *Journal International des Sciences de la Vigne et du Vin, 45(2)*, 61–73.

Lark, R. M., & Marchant, B. P. (2009). Using advanced methods to reduce the cost of soil sampling; ideas for the future. R&D Conference 'Precision in arable farming: current practice and future potential', Grantham, Lincolnshire, UK, 28th-29th October 2009. pp. 18–25 ref.6. https://www.cabdirect.org/cabdirect/FullTextPDF/2010/20103068758.pdf

Martínez-Casasnovas, J.A., Agelet-Fernandez, J., Arnó, J., & Ramos, M.C. (2012). Analysis of vineyard differential management zones and relation to vine development, grape maturity and quality. *Spanish Journal of Agricultural Research, 10(2)*, 326–337.

Meyers, J. M., & Vanden Heuvel, J. E. (2014). Research Note. Use of Normalized Difference Vegetation Index Images to Optimize Vineyard Sampling Protocols. *American Journal of Enology and Viticulture, 65(2)*, 250-253.

Meyers, J. M., Sacks, G. L., van Es, H. M., & Vanden Heuvel, J. E. (2011). Improving vineyard sampling efficiency via dynamic spatially explicit optimisation. *Australian Journal of Grape and Wine Research, 17*, 306-315.

Miranda, C., & Royo, J. B. (2003). A statistical model to estimate potential yields in peach before bloom. *Journal of the American Society for Horticultural Science*, *128*, 297-301.

Miranda, C., Urretavizcaya, I., Santesteban, L. G., & Royo, J. B. (2015). Sampling stratification using aerial imagery to estimate fruit load and hail damage in nectarine trees. In: J. V. Stafford (Ed.) Precision Agriculture'15, Proceedings of the 10th European Conference on Precision Agriculture, Wageningen, The Netherlands: Wageningen Academic Publishers, pp. 541-546.

Miranda, C., Santesteban, L. G., Urrestarazu, J., Loidi, M., & Royo, J. B. (2018). Sampling stratification using aerial imagery to estimate fruit load in peach tree orchards. *Agriculture*, *8*(6), 78; https://doi.org/10.3390/agriculture8060078.

Monestiez, P., Audergon, J. M., & Habib, R. (1990). Spatial dependences and sampling in a fruit tree: a geostatistical approach. *Institut National de la Recherche Agronomique*, Technical Report No. 163, 30 pp.

Moral, F. J., Terrón, J. M., & Marques da Silva, J. R. (2010). Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. *Soil & Tillage Research, 106*, 335–343.

Ortega-Blu, R., & Molina-Roco, M. (2016). Evaluation of vegetation indices and apparent soil electrical conductivity for site-specific vineyard management in Chile. *Precision Agriculture*, *17*, 434-450.

Peralta, N. R., & Costa, J. L. (2013). Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. *Computers and Electronics in Agriculture*, *99*, 218-226.

Soil Survey Staff. (2014). *Keys to Soil Taxonomy, 12th ed*. Washington, USA: USDA-Natural Resources Conservation Service.

Sudduth, K. A., Kitchen, N. R., Wiebold, W. J., Batchelor, W. D., Bollero, G. A., Bullock, D. G., et al. (2005). Relating apparent electrical conductivity to soil properties across the north-central USA. *Computers and Electronics in Agriculture*, 46, 263-283.

Taylor, J., Tisseyre, B., Bramley, R., & Reid, A. (2005). A comparison of the spatial variability of vineyard yield in European and Australian production systems. In: J. V. Stafford (Ed.) Precision Agriculture'05, Proceedings of the 5th European Conference on

POSTPRINT of the article: Uribeetxebarria, A., Martínez-Casasnovas, J.A, Escolà, A., Rosell-Polo, J.R., Arnó, J.2018. Stratified sampling in fruit orchards using cluster-based ancillary information maps: a comparative analysis to improve yield and quality estimates. *Precis.Agric.* DOI: https://doi.org/10.1007/s11119-018-9619-9

Precision Agriculture, Wageningen, The Netherlands: Wageningen Academic Publishers, pp. 907–914.

Taylor, J., Acevedo-Opazo, C., Ojeda, H., & Tisseyre, B. (2010). Identification and significance of sources of spatial variability in grapevine water status. *Australian Journal of Grape and Wine Research, 16*, 218–226.

Uribeetxebarria, A., Arnó, J., Escolà, A., & Martínez-Casasnovas, J.A. (2018). Apparent electrical conductivity and multivariate analysis of soil properties to assess soil constraints in orchards affected by previous parcelling. *Geoderma*, *319*, 185–193.

Vanden Heuvel, J. E., Leonardos, E. D., Proctor, J. T. A., Fisher, K. H., & Sullivan, J. A. (2002). Translocation and partitioning patterns of 14C photoassimilate from light- and shade- adapted shoots in greenhouse-grown 'Chardonnay' grapevines (Vitis vinifera L.). *Journal of the American Society for Horticultural Science, 127*, 912–918.

Webster, R., & Lark, R. M. (2013). *Field sampling for environmental science and management*. London and New York: Routledge, 192 pp.

Wulfsohn, D., Aravena Zamora, F., Potin Téllez, C., Zamora Lagos, I., & García-Fiñana, M. (2012). Multilevel systematic sampling to estimate total fruit number for yield forecasts. *Precision Agriculture*, *13*, 256-275.

Zude-Sasse, M., Fountas, S., Gemtos, T. A., Abu-Khalaf, N. (2016). Applications of precision agriculture in horticultural crops. *European Journal of Horticultural Science, 81(2)*, 78-90.